

From compression to compressed sensing

Shirin Jalali and Arian Maleki

Abstract—Can compression algorithms be employed for recovering signals from their underdetermined set of linear measurements? Addressing this question is the first step towards applying compression algorithms for compressed sensing (CS). In this paper, we consider a family of compression algorithms \mathcal{C}_R , parametrized by rate R , for a compact class of signals $\mathcal{Q} \subset \mathbb{R}^n$. The set of natural images and JPEG2000 at different rates are examples of \mathcal{Q} and \mathcal{C}_R , respectively. We establish a connection between the rate-distortion performance of \mathcal{C}_R , and the number of linear measurement required for successful recovery in CS. We then propose compressible signal pursuit (CSP) algorithm and prove that, with high probability, it accurately and robustly recovers signals from an underdetermined set of linear measurements. We also explore the performance of CSP in the recovery of infinite dimensional signals. Exploring approximations or simplifications of CSP, which is computationally demanding, is left for the future research.

I. INTRODUCTION

The field of compressed sensing (CS) was established on a keen observation that if a signal is sparse in a certain basis it can be recovered from far fewer random linear measurements than its ambient dimension [1], [2]. In the last decade, CS recovery algorithms have evolved to capture more complicated signal structures such as group sparsity, atomic structure, and nuclear norm minimization [3]–[19]. In this paper, we consider a different type of structure based on compression algorithms. Suppose that a class of signals can be “efficiently” compressed by a compression algorithm. Intuitively speaking, such classes of signals have a certain “structure” that enables the compression algorithm to represent them with fewer bits. These structures are often much more complicated than sparsity, and employing them in CS can potentially reduce the number of measurements required for signal recovery.

In this paper, we aim to address the following problem. Is it possible to employ compression schemes in the CS problem and design algorithms that recover signals, either exactly or with “small error”, from random linear measurements? As we will prove in this paper, the answer to this question is affirmative. We propose a CS recovery algorithm based on exhaustive search over the set of “compressible” signals, that, under certain condition on the rate-distortion function, recovers signals from fewer measurements than their ambient dimension. This result provides the first theoretical basis for using compression algorithms in CS.

The organization of the paper is as follows. Section II reviews the main concepts used in this paper and formally states the problem addressed in the paper. Section III summarizes our main contributions. Section IV extends our results to the class of analog signals. Section V reviews the related work in the literature. Section VI includes the proofs of our main theorems. Finally, Section VII concludes the paper.

II. BACKGROUND AND PROBLEM DEFINITION

In this section we first review the concept of compression and rate-distortion function. Then we state the problem we address in this paper more formally.

A. Notation

Boldfaced letters such as \mathbf{x} and \mathbf{X} represent vectors. Calligraphic letters denote sets. Given a finite set \mathcal{A} , $|\mathcal{A}|$ denotes its size. The ℓ_p -norm of $\mathbf{x} \in \mathbb{R}^n$ is defined as $\|\mathbf{x}\|_p \triangleq (\sum_{i=1}^n |x_i|^p)^{1/p}$. The ℓ_0 -norm is also defined as $\|\mathbf{x}\|_0 \triangleq |\{i : x_i \neq 0\}|$. Note that for $p < 1$, $\|\cdot\|_p$ is a semi-norm since it does not satisfy the triangle inequality.

B. Rate-distortion function

Let \mathcal{Q} denote a compact subset of \mathbb{R}^n . Consider a compression algorithm for \mathcal{Q} described by encoder and decoder mappings $(\mathcal{E}, \mathcal{D})$. Encoder

$$\mathcal{E} : \mathcal{Q} \rightarrow \{1, 2, \dots, 2^R\},$$

maps signal $\mathbf{x} \in \mathcal{Q}$ to codeword $\mathcal{E}(\mathbf{x})$. Decoder

$$\mathcal{D} : \{1, 2, \dots, 2^R\} \rightarrow \hat{\mathcal{Q}},$$

maps the coded signal $\mathcal{E}(\mathbf{x})$ back to the reconstruction domain $\hat{\mathcal{Q}}$. Let $\hat{\mathbf{x}} \triangleq \mathcal{D}(\mathcal{E}(\mathbf{x}))$ denote the reconstruction of signal $\mathbf{x} \in \mathcal{Q}$. The performance of the described coding scheme at rate R is measured in terms of its induced distortion defined as

$$D(R) \triangleq \sup_{\mathbf{x} \in \mathcal{Q}} \|\mathbf{x} - \mathcal{D}(\mathcal{E}(\mathbf{x}))\|_2.$$

Throughout the paper, we usually consider a family of fixed-rate compression algorithms $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$ parametrized by the rate R , and its corresponding rate-distortion function $R(D)$ defined as

$$R(D) \triangleq \inf\{R : D(R) \leq D\}.$$

Given compression algorithm $(\mathcal{E}_R, \mathcal{D}_R)$, let \mathcal{C}_R denote its *codebook* defined as

$$\mathcal{C}_R \triangleq \{\mathcal{D}_R(\mathcal{E}_R(\mathbf{x})) : \mathbf{x} \in \mathcal{Q}\}.$$

Note that $|\mathcal{C}_R| \leq 2^R$.

In the remaining of this section we illustrate the concepts with two examples. Let $\mathcal{B}_p^n(\rho) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq \rho\}$ represent a ball of radius ρ in \mathbb{R}^n . Also, let Γ_k^n denote the set of all k -sparse signals in \mathbb{R}^n , i.e.,

$$\Gamma_k^n \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq k\}.$$

Example 1: There exists a family of compression algorithms that achieves the following rate-distortion function on $\mathcal{B}_2^n(\rho)$:

$$R(D) = \frac{1}{2}n \log n + n \log\left(\frac{\rho}{D}\right) + cn,$$

for $D \leq \sqrt{n}$. Here, c is a constant less than 3.

Example 2: There exists a family of compression algorithms that achieves the following rate-distortion function on $\mathcal{B}_2^n(\rho) \cap \Gamma_k^n$:

$$R(D) = \log \binom{n}{k} + k \log \left(\frac{\sqrt{k}\rho}{D} \right) + ck,$$

where c is a constant less than 3.

These are classic examples in the literature. However, we review their proofs in Section VI-B to clarify the concepts introduced here. Note that since $\mathcal{B}_2^n(\rho) \cap \Gamma_k^n \subset \mathcal{B}_2^n(\rho)$, we expect to compress $\mathcal{B}_2^n(\rho) \cap \Gamma_k^n$ more efficiently. This is specially clear as $D \downarrow 0$.

C. Problem statement

Consider the problem of recovering “structured” signal $\mathbf{x} \in \mathbb{R}^n$ from its undersampled set of linear measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{y} \in \mathbb{R}^d$, and $\mathbf{A} \in \mathbb{R}^{d \times n}$, $d < n$, denotes the measurement matrix. For various types of structure such as sparsity, it is well-known that \mathbf{x} may be recovered from measurements \mathbf{y} even though $d < n$. In this paper we explore a more elaborate type of structure based on compressibility.

Instead of being structured as sparse, smooth, etc., suppose that the signal belongs to a compact set $\mathcal{Q} \subset \mathbb{R}^n$ and there exists a family of compression algorithms $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$ with rate-distortion function $R(D)$ for signals in \mathcal{Q} . For instance, we can consider the JPEG2000 compression algorithm [20] at different rates for the class of images. This family of compression algorithms might be exploiting the sparsity of the signal in a certain domain or any other type of structure. The actual mechanism by which the algorithm is compressing the signals in \mathcal{Q} is not important for the purpose of this paper. Instead, we are interested in recovering vector $\mathbf{x} \in \mathcal{Q}$ from an undersampled set of linear equations $\mathbf{y} = \mathbf{A}\mathbf{x}$ by employing the compression algorithms $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$. The question is when this is possible, and what it implies on the rate-distortion function $R(D)$. Since for every compact set we can define a family of compression algorithms, it seems that existence of compression algorithms does not necessarily lead to a CS-recovery method. The following lemma confirms this intuition.

Lemma 1: Let $\mathcal{Q} = \mathcal{B}_2^n(1)$. If the number of measurements is less than the ambient dimension n , any CS-recovery algorithm will result in an ℓ_2 reconstruction error of at least 1, for any measurement matrix.

Proof: Consider any reconstruction algorithm for the signals in $\mathcal{B}_2^n(1)$ based on their linear measurements acquired by measurement matrix $\mathbf{A} \in \mathbb{R}^{d \times n}$, with $d < n$. Let $\text{Ker}(\mathbf{A}) \triangleq \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. Since $d < n$, $\text{Ker}(\mathbf{A}) - \{\mathbf{0}\} \neq \emptyset$. All signals in $\text{Ker}(\mathbf{A}) \cap \mathcal{B}_2^n(1)$ are mapped to the all-zero measurement vector, and hence the recovery algorithm maps all of them to some $\hat{\mathbf{x}}_0 \in \mathbb{R}^n$. It is straightforward to confirm that

$$\inf_{\hat{\mathbf{x}}_0} \sup_{\mathbf{x} \in \text{Ker}(\mathbf{A}) \cap \mathcal{B}_2^n(1)} \|\mathbf{x} - \hat{\mathbf{x}}_0\|_2 = 1.$$

In fact the best reconstruction for $\mathbf{x} \in \text{Ker}(\mathbf{A}) \cap \mathcal{B}_2^n(1)$ is $\hat{\mathbf{x}}_0 = \mathbf{0}$, which leads to $\sup_{\mathbf{x} \in \text{Ker}(\mathbf{A}) \cap \mathcal{B}_2^n(1)} \|\mathbf{x} - \hat{\mathbf{x}}_0\|_2 = 1$. ■

Therefore, the first step in employing compression algorithms for CS is to characterize the class of compression algorithms that can potentially lead to CS-recovery methods.

Definition 1: Compressed sensing is said to be *applicable* to a compact set $\mathcal{Q} \subset \mathbb{R}^n$ with $d < n$ measurements, if, for any $\epsilon > 0$, there exists a $d \times n$ matrix \mathbf{A}_ϵ and a recovery algorithm \mathcal{A}_ϵ ,

$$\mathcal{A}_\epsilon : \mathbb{R}^d \rightarrow \mathbb{R}^n,$$

such that $\|\mathcal{A}_\epsilon(\mathbf{A}_\epsilon \mathbf{x}) - \mathbf{x}\|_2 \leq \epsilon$, for all $\mathbf{x} \in \mathcal{Q}$.

According to Lemma 1, CS is not applicable to $\mathcal{B}_2^n(1)$ with d measurements, for any $d < n$. Next we define α -dimension for a rate-distortion function and establish its connection with CS-applicability.

Definition 2: Consider compact set $\mathcal{Q} \subset \mathbb{R}^n$, and a family of fixed-rate compression codes, $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$, with rate-distortion function $R(D)$. Define the high resolution rate distortion dimension, or α -dimension, of a family of codes as

$$\alpha \triangleq \limsup_{D \rightarrow 0} \frac{R(D)}{\log(\frac{1}{D})}. \quad (1)$$

In Section V we discuss the connection between α -dimension and other well-known concepts in information theory and functional analysis.

Consider a compact set $\mathcal{Q} \subset \mathbb{R}^n$, and without loss of generality, assume that $\mathbf{0} \in \mathcal{Q}$. Since the set is compact

$$\rho(\mathcal{Q}) \triangleq \sup_{\mathbf{x} \in \mathcal{Q}} \|\mathbf{x}\|_2$$

is finite. Therefore, $\mathcal{Q} \subset \mathcal{B}_2^n(\rho)$, and according to Example 1, there exists a family of compression algorithms with

$$R(D) \leq n \log \left(\frac{\rho}{D} \right) + c, \quad (2)$$

where $\rho = \rho(\mathcal{Q})$, $D < \rho$, and c is a constant independent of the distortion level D . Therefore, for any compact set \mathcal{Q} , there exists a family of compression codes with α -dimension upper-bounded by n . The interesting regime is when there exists a family of codes with α -dimension strictly smaller than n . For instance, the set of k -sparse signals in $\mathcal{B}_2^n(1)$, discussed in Example 2, is a set for which there exists a family of compression algorithms with α -dimension smaller than n . In the remaining of this section, we explore the connection between the CS-applicability of a compact set \mathcal{Q} and the α -dimension of a family of codes for \mathcal{Q} . The question is whether CS is applicable to \mathcal{Q} with number of measurements $d < n$, and if the answer is affirmative, what is the minimum number of measurements for this result to hold.

Given $\alpha > 0$, let \mathcal{S}_α^n denote the set of all subsets of $\mathcal{B}_2^n(1)$ for which there exists a family of compression algorithms with α -dimension upper-bounded by α . For each $\mathcal{Q} \in \mathcal{S}_\alpha^n$, define $d_{\min}(\mathcal{Q})$ as the minimum number of measurements for which CS is applicable to \mathcal{Q} . The following theorem provides a lower bound for the number of measurements.

Theorem 1: If CS is applicable to any element of $\mathcal{S}_{\alpha,n}$ with d measurements, then $d \geq \lfloor \alpha \rfloor$. In other words,

$$\sup_{\mathcal{Q} \in \mathcal{S}_\alpha^n} d_{\min}(\mathcal{Q}) \geq \lfloor \alpha \rfloor.$$

Proof: Set $k \triangleq \lfloor \alpha \rfloor$ and define Γ_k as the set of vectors in $\mathcal{B}_2^n(1)$, whose $n - k$ final coordinates are equal to zero, i.e.,

$$\Gamma_k \triangleq \{\mathbf{x} \in \mathcal{B}_2^n(1) : x_{k+1} = x_{k+2} = \dots = x_n = 0\}.$$

As shown in Example 1, $\Gamma_k \in \mathcal{S}_\alpha^n$. Also a very simple modification of Lemma 1 proves that if the number of measurements is less than k , then the reconstruction error of any recovery algorithm exceeds 1. Therefore, CS will not be applicable to Γ_k . ■

Note that the notion of CS-applicability with d measurements is the minimal requirement for the practical use of CS. In particular, it does not require the robustness of the algorithm to measurement noise. Also, the measurement matrix can be adapted to the structure of data and the recovery algorithm can exploit any extra information about the set. In the rest of this paper, we show that considering random measurement matrices (nonadaptive measurements) and following Occam's principle results in an accurate and stable recovery algorithm.

Our algorithm searches over the space of "compressible signals" and finds the one that matches the measurements the best. More formally, given a compression algorithm $(\mathcal{E}_R, \mathcal{D}_R)$ with codebook \mathcal{C}_R , consider *compressible signal pursuit* (CSP) algorithm for recovering $\mathbf{x}_o \in \mathcal{Q}$ from its measurements $\mathbf{y}_o = A\mathbf{x}_o$ defined as

$$\hat{\mathbf{x}}_o = \arg \min_{\mathbf{c} \in \mathcal{C}_R} \|\mathbf{y}_o - A\mathbf{c}\|_2^2. \quad (3)$$

Here the rate R can be considered as a free parameter, which, as we will see in Section III, plays a role in the tradeoff between the success probability of CSP, and its reconstruction error. Note that we still ignore one important aspect of practical algorithms and that is "computational complexity". CSP is based on an exhaustive search and hence is computationally very demanding. Practical implementation of such ideas is left for future research.

III. MAIN CONTRIBUTIONS

Consider the problem of recovering signal $\mathbf{x}_o \in \mathcal{Q} \subset \mathbb{R}^n$, from $d < n$ linear measurements $\mathbf{y}_o = A\mathbf{x}_o + \mathbf{z}$, where the entries of A are i.i.d. $\mathcal{N}(0, 1)$, and $\mathbf{z} \in \mathbb{R}^d$ represents the measurements noise in the system. Furthermore, assume that there exists a family of compression algorithms, $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$, for the signals of \mathcal{Q} , which has rate-distortion function $R(D)$. We employ the CSP algorithm described in (3) to recover \mathbf{x}_o from \mathbf{y}_o .

A. Noiseless measurements

Our first result is concerned with the performance of the CSP algorithm, when there is no noise in the system, i.e., $\mathbf{z} = \mathbf{0}$.

Theorem 2: Consider compression code $(\mathcal{E}, \mathcal{D})$ for set \mathcal{Q} operating at rate R and distortion D . Let $A \in \mathbb{R}^{d \times n}$, where $A_{i,j}$ are i.i.d. $\mathcal{N}(0, 1)$. For $\mathbf{x}_o \in \mathcal{Q}$, let $\hat{\mathbf{x}}_o$ denote the reconstruction of \mathbf{x}_o from $\mathbf{y}_o = A\mathbf{x}_o$, $A \in \mathbb{R}^{d \times n}$, by the CSP algorithm employing $(\mathcal{E}, \mathcal{D})$. Then,

$$\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \leq D \sqrt{\frac{1 + \tau_1}{1 - \tau_2}},$$

with probability at least

$$1 - 2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))} - e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))}.$$

Theorem 2 proves that in many cases CSP algorithm provides an "accurate estimate" of \mathbf{x}_o with a number of measurements that is less than the ambient dimension n . Corollary 1 below describes one instance of such cases.

Corollary 1: Let $D < e^{-1}$ and $d \geq \frac{4R(D)}{\log(1/eD)}$. Then

$$P(\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \geq \sqrt{2D}) \leq e^{-R(D)} + e^{-\frac{R(D)}{8 \log(1/eD)}}.$$

Proof:

Let $\tau_2 = 1 - D$ in Theorem 2. Then,

$$D \sqrt{\frac{1 + \tau_1}{1 - \tau_2}} = D \sqrt{\frac{1 + \tau_1}{D}} \leq \sqrt{2D}. \quad (4)$$

If $d \geq \frac{4R(D)}{\log(1/eD)}$, then

$$\begin{aligned} & R(D) \log 2 + \frac{d}{2}(\tau_2 + \log(1 - \tau_2)) \\ &= R(D) \log 2 + \frac{d}{2}(1 - D + \log D) \\ &\leq R(D) \log 2 + \frac{2R(D)}{\log(1/eD)}(1 - D + \log D) \\ &\leq R(D) \log 2 + \frac{2R(D)}{\log(1/eD)} \log(eD) \\ &\leq -R(D). \end{aligned} \quad (5)$$

Hence,

$$2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))} \leq e^{-R(D)}. \quad (6)$$

On the other hand, for $\tau_1 \leq 1$,

$$\tau_1 - \log(1 + \tau_1) \geq \tau_1^2/4.$$

Therefore,

$$\frac{d}{2}(\tau_1 - \log(1 + \tau_1)) \geq \frac{R(D)\tau_1^2}{2 \log(1/eD)}.$$

Set $\tau_1 = 0.5$. Then,

$$e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))} \leq e^{-\frac{R(D)}{8 \log(1/eD)}}. \quad (7)$$

The desired result follows from combining the bounds in (4), (6) and (7). ■

Note that if $D \ll 1$ then $\frac{R(D)}{\log(1/eD)}$ is much smaller than $R(D)$ itself. In fact, according to (2) as $D \rightarrow 0$, $\limsup_{D \rightarrow 0} \frac{R(D)}{\log(1/eD)} \leq n$. The following corollary characterizes the number of measurements required and probability of correct recovery as a function of α -dimension.

Corollary 2: Consider a family of rate-distortion codes $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$ with α -dimension α . Then for every $\epsilon > 0$, there exists $R > 0$ and a corresponding code $(\mathcal{E}_R, \mathcal{D}_R)$, such that if we employ this code in the CSP algorithm with

$$d > 4\alpha$$

measurements, then

$$P(\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \geq \epsilon) \leq e^{-0.1\alpha}.$$

Corollary 2 directly follows from Corollary 1, by taking the limit as $D \rightarrow 0$. Example 2 shows an application of this corollary.

Example 3: Consider the class of k -sparse signals in \mathbb{R}^n discussed in Example 2. It is straightforward to check that there exists a family of codes such that $R(D)/\log(1/D) \downarrow k$ as $D \downarrow 0$. Therefore, from Corollary 2, for every $\epsilon > 0$, the solution of the CSP algorithm satisfies

$$P(\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \geq \epsilon) \leq e^{-0.1k},$$

if $d > 4k$.

By choosing different values for τ_1 and τ_2 we can derive different upper bounds for the recovery error. Here is another instance of such result.

Corollary 3: Taking the number of measurements d such that $d \geq 4R(D)/\log n$, then

$$P\left(\frac{1}{\sqrt{n}}\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \leq \sqrt{2}D\right) \leq e^{-R(D)} + e^{-\frac{R(D)}{8\log n}}.$$

Proof: Choose $\tau_2 = 1 - 1/n$ in Theorem 2. Then, if $d \geq 4R(D)/\log n$ and n is large enough,

$$\begin{aligned} R(D) \log 2 + \frac{d}{2}(\tau_2 + \log(1 - \tau_2)) \\ = R(D) \log 2 + \frac{2R(D)}{\log n}(1 - n^{-1} - \log n) \\ \leq -R(D). \end{aligned}$$

Also, choosing $\tau_1 = 0.5$, it follows that

$$\frac{d}{2}(\tau_1 - \log(1 + \tau_1)) \geq \frac{R(D)}{8\log n},$$

and

$$D\sqrt{\frac{1 + \tau_1}{1 - \tau_2}} \leq \sqrt{2n}D. \quad (8)$$

As a final remark, we derive a lower bound for the number of measurements required according to Theorem 1 for the “success” of the CSP algorithm. Consider the success probability in Theorem 1. To keep success probability larger than zero we require

$$R \log(2) + \frac{d}{2}(\tau_1 - \log(1 - \tau_1)) < 0 \quad (9)$$

Therefore, to reduce the number of measurement, we require τ_1 to be large. But τ_1 is always less than 1. Furthermore, if $\tau_1 > 1 - D^2$, the upper bound on the reconstruction error will be larger than 1, which is a trivial bound for any signal in $B_2^n(1)$. Hence, we consider $\tau_1 < 1 - D^2$. If we set $\tau_1 = 1 - D^2$ in (9), we obtain $d \geq \frac{2\log(2)R(D)}{2\log(D)+1-D^2}$.

B. Noisy measurements

An inevitable part of any measurement system is noise. In this section we analyze the performance of CSP in the presence of noise. Consider the case where the linear measurements are corrupted by i.i.d. noise, i.e., $\mathbf{y}_o = A\mathbf{x}_o + \mathbf{z}$, where $z_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, d$. To recover signal \mathbf{x}_o from \mathbf{y}_o , again we employ the CSP algorithm described in (3).

Theorem 3: Let $\hat{\mathbf{x}}_o$ denote the solution of CSP to input \mathbf{y}_o , using fixed-rate code $(\mathcal{E}, \mathcal{D})$ for compact set \mathcal{Q} , which operates at rate R and distortion $D \leq (5e)^{-1}$. For any $\mathbf{x}_o \in \mathcal{Q}$ and $\eta > 1$, choosing

$$d = \left\lceil \frac{4\eta R}{\log \frac{1}{eD}} \right\rceil,$$

then

$$\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \leq \frac{\sigma}{D\sqrt{\eta}} \log \frac{1}{eD} + \sqrt{2D + \frac{2\sigma}{\sqrt{\eta}} + \frac{\sigma^2}{\eta D^2} \log^2 \left(\frac{1}{eD} \right)},$$

with probability exceeding

$$\begin{aligned} 1 - e^{-\frac{R}{\log 1/(eD)}} - 2e^{-\frac{0.6\eta R}{\log 1/(eD)}} - e^{-(2\eta-1)R} \\ - e^{-(0.8\eta-\log 2)R} - e^{-0.3R}. \end{aligned} \quad (10)$$

Note that D (or equivalently R) acts as the free parameter of the CSP algorithm and control the bias and variance of the final estimate. Intuitively speaking small values of D lead to large variance since σ is divided by D in two terms. Large values of D make the variance small, but increase the bias (due to $2D$ term under the radical sign). The optimal choice of the free parameter R is dictated by the rate-distortion performance of the code, number of measurements, and the variance of the noise.

IV. EXTENSION TO ANALOG SIGNALS

So far, we have considered the problem of recovering finite-dimensional signals from their undersampled set of linear measurements. But the framework we have developed is applicable to infinite-dimensional spaces as well. In this section, we extend our results to recovering continuous-time function $f : [0, 1] \rightarrow \mathbb{R}$ from a finite number of random linear measurements. In this section, we first review the related basic concepts required for analyzing continuous-time functions and then propose a recovery algorithm similar to CSP for such signals.

A. Ito's integral

For continuous-time signals, we consider a measurement that is based on the Wiener process. Wiener process $W(t)$, a.k.a. Brownian motion, is a continuous time process that satisfies the following four properties:

- 1) $P(W_i(0) = 0) = 1$.
- 2) The probability that a randomly generated path to be continuous is equal to 1.
- 3) $W_i(t) - W_i(s) = N(0, t - s)$, for $0 \leq s < t \leq 1$.
- 4) For $0 \leq s_1 \leq t_1 \leq s_2 \leq t_2$, $W_{t_1} - W_{s_1}$ is independent of $W_{t_2} - W_{s_2}$.

This process is a key component of stochastic calculus and stochastic differential equations. In particular, Ito's integral, which plays a central role in stochastic differential equations, is defined based on the Wiener process. To keep our discussions simple, we introduce a specific form of the Ito's integral that is used in this paper.

For function $f : [0, 1] \rightarrow \mathbb{R}$, define its p -norm as

$$\|f\|_p \triangleq \left(\int_0^1 |f(t)|^p dt \right)^{1/p}.$$

Furthermore, define $L_p([0, 1])$ as the set of functions from $[0, 1]$ to \mathbb{R} with finite p norm, i.e.,

$$L_p([0, 1]) \triangleq \{f : [0, 1] \rightarrow \mathbb{R} \mid \|f\|_p < \infty\}.$$

In this paper we are mainly interested in $L_2([0, 1])$, which includes the set of functions with finite second moment. Also, for some technicalities that become clear later on, we restrict our attention to subset $L_2^r([0, 1])$ of $L_2([0, 1])$ that is defined as the set of function in $L_2([0, 1])$ that are piecewise continuous with a finite number of discontinuities.

Suppose that $f_s \in L_2^r([0, 1])$ is a simple function, i.e., f_s can be represented as

$$f_s(t) = \sum_{k=1}^N c_k \mathbb{1}_{t \in (t_k, t_{k+1}]},$$

where $0 = t_1 < t_2 < \dots < t_N = 1$, and $(c_1, \dots, c_N) \in \mathbb{R}^N$. For such functions, Ito's stochastic integral is defined as

$$\int_0^1 f_s(t) dW(t) \triangleq \sum_{k=1}^N c_k (W(t_{k+1}) - W(t_k)).$$

Note that since $(W(t_{i+1}) - W(t_i) : i = 0, 1, \dots, N-1)$ are independent Gaussian random variables, the result of this integral is a Gaussian random variable with mean zero and variance $\sum_{k=1}^N c_k^2 (t_{k+1} - t_k)$. For $f \in L_2^r([0, 1])$, let (f_1, f_2, \dots) be a sequence of simple functions such that

$$\lim_{n \rightarrow \infty} \int_0^1 (f(t) - f_n(t))^2 dt = 0.$$

Then the Ito's integral of f is defined as

$$\int_0^1 f(t) dW(t) \triangleq \lim_{n \rightarrow \infty} \int_0^1 f_n(t) dW(t),$$

where the convergence is in the mean square sense. As we will prove in Section VI-E Ito's integral is defined for any function in $L_2^r([0, 1])$. We will also show that the result of the integral is a Gaussian random variable.

B. Rate-distortion function

Consider a class of functions $\mathcal{F} \subset L_2^r([0, 1])$, and a family of compression algorithms $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$, indexed by rate R . For each code in this family, the encoder and decoder mappings, $(\mathcal{E}_R, \mathcal{D}_R)$, are defined as

$$\mathcal{E}_R : \mathcal{F} \rightarrow \{1, 2, \dots, 2^R\},$$

and

$$\mathcal{D}_R : \{1, 2, \dots, 2^R\} \rightarrow \hat{\mathcal{F}},$$

respectively, where $\hat{\mathcal{F}} \subset L_2([0, 1])$ denotes the class of reconstruction functions. For a function $f \in \mathcal{F}$, $\mathcal{D}_R(\mathcal{E}_R(f))$ denotes the reconstruction of function f by the code $(\mathcal{E}_R, \mathcal{D}_R)$. Given compression algorithm $(\mathcal{E}_R, \mathcal{D}_R)$, let \mathcal{C}_R denote its codebook defined as

$$\mathcal{C}_R \triangleq \{\mathcal{D}_R(\mathcal{E}_R(f)) : f \in \mathcal{F}\}.$$

The distortion-rate function of this family of codes is defined as

$$D(R) \triangleq \sup_{f \in \mathcal{F}} \|f - \mathcal{D}_R(\mathcal{E}_R(f))\|_2.$$

C. Compressed sensing of analog signals

1) *Measurement process*: Unlike the classical compressed sensing setup, where the measurement process is assumed to be in the discrete time domain, here we consider analog domain measurements. In particular, for function $f_o \in L_2^r([0, 1])$, we consider d linear measurements of the form

$$y_i = \int_0^1 f_o(t) dW_i(t), \quad \text{for } i = 1, 2, \dots, d, \quad (11)$$

where W_i , $i = 1, 2, \dots, d$, are independent Wiener processes. Similar to the discrete time settings, each measurement is a random linear combination of the signal at different times. As we will show in this section, this type of measurement process ensures that with "sufficient" number of measurement, the "critical information" about the signal is acquired by the measurements, and therefore, we can recover f_o from the measurement vector $\mathbf{y} \in \mathbb{R}^d$.

2) *CSP algorithm*: Consider $\{(\mathcal{E}_R, \mathcal{D}_R) : R > 0\}$ a family of compression algorithms for class of functions $\mathcal{F} \subset L_2^r([0, 1])$ with rate-distortion function $D(R)$. We are interested in recovering a function $f_o \in \mathcal{F} \subset L_2^r([0, 1])$ from d linear measurements, $\mathbf{y}_o \in \mathbb{R}^n$ such that, for $i = 1, \dots, d$,

$$y_{o,i} \triangleq \int_0^1 f_o(t) dW_i(t). \quad (12)$$

Let $\mathcal{A} : L_2([0, 1]) \rightarrow \mathbb{R}^d$ denote the just-defined linear measurement process, i.e., $\mathbf{y}_o = \mathcal{A}(f_o)$. To recover the function f_o from \mathbf{y}_o , we employ the CSP algorithm defined as

$$\hat{f}_o = \arg \max_{\hat{f} \in \mathcal{C}_R} \|\mathbf{y}_o - \mathcal{A}(\hat{f})\|_2^2. \quad (13)$$

The intuition for the CSP algorithm is the same as what we proposed before; among all the low-complexity signals (defined according to the compression algorithm) look for the one that matches the measurements the best. The parameter R can be considered as a free parameter in the algorithm, whose role will be clear in the next section.

3) *Performance guarantees for CSP*: Consider the problem of recovering function $f_o \in \mathcal{F} \subset L_2^r([0, 1])$ from its under-sampled set of d random linear measurements, $\mathbf{y}_o = \mathcal{A}(f_o)$, as defined in (12). Assume that there exists a family of compression algorithms for \mathcal{F} indexed by R , $(\mathcal{E}, \mathcal{D})$, that achieves the rate distortion function $R(D)$. We employ the CSP algorithm to recover f_o . The following theorem characterizes the performance of the CSP algorithm.

Theorem 4: For $f_o \in \mathcal{F}$, let \hat{f}_o denote the reconstruction of f_o from $\mathbf{y}_o = \mathcal{A}(f_o)$, by the CSP algorithm employing rate- R compression algorithm $(\mathcal{E}, \mathcal{D})$. Then,

$$\|\hat{f}_o - f_o\|_2 \leq D \sqrt{\frac{1 + \tau_1}{1 - \tau_2}},$$

with probability at least

$$1 - 2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))} - e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))}.$$

See Section VI-E for the proof.

Theorem 4 considers noiseless measurements. In the case of noisy measurements, assume that

$$y_{o,i} = \int_0^1 f_o(t) dW_i(t) + z_i,$$

where $z_i \sim N(0, \sigma^2)$ represents the measurement noise. As the following result shows, even in this infinite-dimensional setting, the algorithm is robust to noise.

Theorem 5: Let \hat{f}_o denote the solution of CSP to input \mathbf{y}_o , using code $(\mathcal{E}, \mathcal{D})$ at rate R and distortion D for compact set \mathcal{F} . For distortion $D \leq (5e)^{-1}$, and for any $f_o \in \mathcal{Q}$ and $\eta > 1$, choosing

$$d = \left\lceil \frac{4\eta R}{\log \frac{1}{eD}} \right\rceil,$$

then

$$\|\hat{f}_o - f_o\|_2 \leq \frac{\sigma}{D\sqrt{\eta}} \log \frac{1}{eD} + \sqrt{2D + \frac{2\sigma}{\sqrt{\eta}} + \frac{\sigma^2}{\eta D^2} \log^2\left(\frac{1}{eD}\right)},$$

with probability exceeding

$$1 - e^{-\frac{R}{\log 1/(eD)}} - 2e^{-\frac{0.6\eta R}{\log 1/(eD)}} - e^{-(2\eta-1)R} - e^{-(0.8\eta-\log 2)R} - e^{-0.3R}. \quad (14)$$

After employing Lemmas 4 and 5, the proof of this result is similar to the proof of Theorem 3. Hence we skip the proof.

Note that according Comparing Theorems 2 and 3 with Theorem 4 and 5 may lead us to a conclusion that the ambient dimension of the signal is not important in the performance of CSP algorithm. However, this conclusion is not true. For further information regarding this issue see Section IV-E.

D. Applications

In this section, we investigate the implications of Theorem 4 for three different classes of continuous time signals. As we will see in these examples, in case of infinite dimensional signals, the rate-distortion performance shows more diverse types of behavior. Different rate distortion behaviors of these classes clarify the opportunities and limitations of analog CS.

Let $\mathcal{P}_N^Q(A)$ denote the class of piecewise polynomial functions with N , Q , A representing the maximum degree of the polynomials, number of singularity points¹, and maximum value of the function, respectively.

Example 4: There exists a family of compression algorithms for $\mathcal{P}_N^Q(A)$ that achieves

$$R(D) = (N + 3)(Q + 1) \log \left(\frac{1}{D} \right) + c,$$

where c is a constant independent of D [21].

The rate-distortion behavior described in Example 4 for $\mathcal{P}_N^Q(A)$ is reminiscent of the rate-distortion function of finite-

dimensional spaces. However, since the locations of singularities are not fixed, the space is not finite dimensional. Nevertheless, we would expect to recover the signals of this class with finite number of measurements with small error.

Corollary 4: For every $\epsilon > 0$, there exists $R > 0$ and a corresponding code $(\mathcal{E}_R, \mathcal{D}_R)$ at rate R , such that if we employ this code in the CSP algorithm with

$$d > 4(N + 3)(Q + 1)$$

measurements, then for any $f_o \in \mathcal{P}_N^Q(A)$, the reconstruction error satisfies

$$P(\|\hat{f}_o - f_o\|_2 \geq \epsilon) \leq e^{-0.1(N+3)(Q+1)}.$$

Due to the similarity of Theorem 4 and Theorem 2, the proof of this corollary is exactly the same as the proof of Lemma 1.

For finite-dimensional spaces we could show that for any compact subset of \mathbb{R}^n there exists a compression algorithm whose rate distortion function satisfies $R(D) = O(n \log(\frac{1}{D}))$. However, in infinite dimensional spaces this is not the case any more. The next example illustrates a class with a slightly different rate-distortion behavior.

Let $\mathcal{H}_h(C)$ be a class of functions $f : \mathbb{C} \rightarrow \mathbb{C}$ satisfying the following properties:

- a) f is analytic on a strip of size h , i.e., $f(z)$ is analytic on $\{z = x + iy \mid |y| \leq h\}$.
- b) $|f(z)|$ is bounded by C .

Define $\mathcal{G}_h(C)$ as

$$\mathcal{G}_h(C) \triangleq \{g : [0, 1] \rightarrow \mathbb{R} : \exists f \in \mathcal{H}_h(C), g(x) = |f(x + i0)|\}.$$

Example 5: There exists a family of compression algorithms for $\mathcal{G}_h(C)$ that achieves

$$R(D) \leq c \left(\log \frac{1}{D} \right)^2,$$

where c is a constant that does not depend on D [22].

Clearly for this class of functions the α -dimension is infinite, therefore we do not expect results similar to Corollary 4 to hold for this class. However CSP algorithm is still useful for this class as is described in the next corollary.

Corollary 5: Let $D \ll 1$ and $d \geq 4c \log(1/D)$. Then

$$P(\|\hat{\mathbf{x}}_o - \mathbf{x}_o\|_2 \geq \sqrt{2D}) \leq e^{-c \log^2(1/D)} + e^{-\frac{c \log(1/D)}{8}}.$$

The proof is similar to the proof of Corollary 1.

While the number of required measurements tends to infinity, as the distortion goes to zero, it only grows logarithmically with the distortion. Therefore, intuitively speaking, accurate estimates are still obtained from few measurements.

If the class of functions is too rich (less structured), then the growth rate of rate-distortion will be faster and therefore to obtain reasonably accurate reconstruction we may require many observations. Here, we present one such example.

Consider the class of smooth functions for which the coders require higher bit-rates to achieve the same distortion level. Let $H^\alpha(C)$ be the class of real functions $f : [0, 1] \rightarrow \mathbb{R}$, having derivative of order α in L^2 (in the sense of Riemann-Liouville) uniformly bounded by some constant C .

¹Singularity point is a point at which the signal is not infinitely differentiable

Example 6: There exists a family of compression algorithms for $H^\alpha(C)$ that achieves

$$R(D) = c \left(\frac{1}{D} \right)^{\frac{1}{\alpha}},$$

where c is a constant independent of D [22].

According to Theorem 4, having $O(\frac{1}{D^{1/\alpha} \log(1/D)})$ measurements, the reconstruction error is bounded by $\sqrt{2D}$. But $O(\frac{1}{D^{1/\alpha} \log(1/D)})$ is much larger than the number of bits that are required for achieving distortion $\sqrt{2D}$, i.e., $O(\frac{1}{D^{1/(2\alpha)}})$. Therefore, in such cases, CSP algorithm is not particularly interesting. In fact the number of measurements required grows rapidly as we decrease the distortion compared to other classes.

E. Discussion

The results we have discussed so far are the same for finite and infinite dimensional classes. Such results may mislead us to a conclusion that as long as the performance of CSP is concerned the ambient dimension is not important. However, the finiteness of ambient dimension may help derive stronger results. Next theorem is an instance of such results.

Theorem 6: Let $A \in \mathbb{R}^{d \times n}$ be a measurement matrix with $A_{ij} \stackrel{iid}{\sim} N(0, 1)$. For any $\mathbf{x}_o \in \mathcal{Q}$, we denote the reconstruction of the CSP algorithm at rate R with $\hat{\mathbf{x}}_o$. We have

$$\begin{aligned} & \mathbb{P} \left(\forall \mathbf{x}_o \in \mathcal{Q} : \|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \geq \frac{2}{1-\tau} \left(\sqrt{\frac{n}{d}} + (1+t) \right) D \right) \\ & \leq e^{-dt^2/2} + 2^{2R} e^{\frac{d}{2}(\tau + \log(1-\tau))}. \end{aligned}$$

See Section VI-F for the proof.

Note that there is a major difference between Theorem 6 and Theorem 2. Theorem 6 claims that once we draw a random matrix from Gaussian distribution, this matrix with high probability works for any signal in \mathcal{Q} . However, Theorem 2 considers individual sequences. Note that the strength of Theorem 6 has come at a price of larger reconstruction error and lower success probability.

V. RELATED WORK

A. Connection of compression and compressed sensing

In this paper we consider the problem of using a family of compression algorithms for compressed sensing. The other direction, i.e., using CS for compression have also been extensively studied in the literature [23]–[31]. In this line of work the rate-distortion that is achieved by scalar (or in a few cases adaptive) quantization of random linear measurements has been derived. However, such results are different from our work since they only consider either sparse or approximately sparse signals. Furthermore, we consider a different direction, that is, the direction of deriving CS recovery algorithms based on compression schemes.

B. Kolmogorov's ϵ -entropy and compressed sensing

The ϵ -entropy of a compact set \mathcal{Q} is defined as

$$H_\epsilon(\mathcal{Q}) = \log_2 N_\epsilon(\mathcal{Q}),$$

where $N_\epsilon(\mathcal{Q})$ is the minimum number of elements in an ϵ -covering of \mathcal{Q} [22]. The ϵ -entropy, H_ϵ , provides a lower bound on the rate distortion of any family of compression algorithms. In other words, if $R(D)$ is the rate distortion function of a family of compression algorithms on \mathcal{Q} , then

$$R(D) \geq H_D(\mathcal{Q}).$$

It is clear that our results can be stated in terms of Kolmogorov's ϵ -entropy by considering it as the optimal compression scheme from the perspective of rate-distortion tradeoff. In particular, Corollary 2 leads to the following result:

Corollary 6: Consider compact set $\mathcal{Q} \subset \mathbb{R}^n$. For $\epsilon > 0$, let \mathcal{C}_ϵ denote an ϵ -covering of $\mathcal{Q} \subset \mathbb{R}^n$ such that $\log |\mathcal{C}_\epsilon| = H_\epsilon(\mathcal{Q})$, and assume that

$$\limsup_{\epsilon \rightarrow 0} \frac{H_\epsilon(\mathcal{Q})}{\log(1/\epsilon)} \leq \alpha.$$

For $\mathbf{x}_o \in \mathcal{Q}$, and $\epsilon > 0$, let $\mathbf{x}_{o,\epsilon}$ denote the reconstruction of CSP employing \mathcal{C}_ϵ , i.e.,

$$\hat{\mathbf{x}}_{o,\epsilon} = \arg \min_{\mathbf{c} \in \mathcal{C}_\epsilon} \|\mathbf{A}\mathbf{c} - \mathbf{A}\mathbf{x}_o\|.$$

If $d > 4\alpha$, then, choosing ϵ small enough,

$$\mathbb{P}(\|\hat{\mathbf{x}}_{o,\epsilon} - \mathbf{x}_o\|_2 \geq \sqrt{2\epsilon}) \leq e^{-0.1\alpha}.$$

The quantity

$$\limsup_{\epsilon \rightarrow 0} \frac{H_\epsilon(\mathcal{Q})}{\log(1/\epsilon)},$$

called upper metric dimension [22] or Minkowski dimension [32]. Metric dimension is a measure of the massiveness of compact sets in finite dimensional spaces [22]. The connection between Minkowski dimension and CS has also been explored in the stochastic settings that will be reviewed in Section V-C.

C. Stochastic settings

This paper considers a deterministic signal model. However, stochastic settings have also been considered in CS [33]–[43]. In such models the data is assumed to follow a certain distribution (often i.i.d.) and the probability of correct recovery is measured as the ambient dimension tends to infinity. In many cases the algorithms exhibit certain phase transitions in the probability of correct recovery. Such phase transitions have been characterized in certain cases either theoretically or empirically [37], [38], [40]–[42].

The most relevant to our work are [40], [42]. These two papers characterize the performance of “information-theoretically” optimal algorithms in the asymptotic setting. For instance they prove that the number of measurements that are required for “exact” recovery is the same as the Rényi information dimension. Even though there is an interesting connection between Rényi information dimension and metric dimension [44], there are several major differences between our work and the work of [40], [42]. First our framework is concerned with the deterministic signal models. Second, our results are for finite-dimensional signals, and are non-asymptotic. Third, we consider arbitrary family of compression algorithms and characterizes when such schemes can be used for signal recovery from random linear measurements.

D. Kolmogorov complexity

Our work is mainly inspired by series of work on the connection between Kolmogorov complexity of sequences and CS [45]–[51]. In particular, [45] defines the *Kolmogorov information dimension* of $\mathbf{x} = (x_1, x_2, \dots, x_n) \in [0, 1]^n$ at resolution m as

$$\kappa_{m,n}(\mathbf{x}) \triangleq \frac{K^{[\cdot]^m}(x_1, x_2, \dots, x_n)}{m},$$

where intuitively speaking, $K^{[\cdot]^m}(x_1, x_2, \dots, x_n)$ denotes the Kolmogorov complexity of vector \mathbf{x} where each component is quantized by m bits, and proves that if the Kolmogorov information dimension of a sequence is small compared to its ambient dimension one can recover it from an undersampled set of linear measurements. Our results have several connections with [45]. The proof techniques we use here have similarities to the proof techniques used in [45]. However, not only the problems are different, but also there are some major differences. One example is the continuous time domain that requires different treatment of the problem. Also, our result is the first step in a new direction toward practical implementation of [45]. While the CSP algorithm is computationally demanding at this point, it provides an approach to designing sub-optimal algorithms such as greedy methods. Furthermore, CSP algorithm may enable us to employ universal compression algorithms [52], [53] and develop universal compressed sensing methods. This has been the main goal of [45], [47]–[51]

VI. PROOFS

A. Background

We use the following two lemmas from [51] throughout our proofs.

Lemma 2 (χ^2 -concentration): Fix $\tau > 0$, and let $Z_i \sim \mathcal{N}(0, 1)$, $i = 1, 2, \dots, d$. Then,

$$\mathbb{P}\left(\sum_{i=1}^d Z_i^2 < d(1 - \tau)\right) \leq e^{\frac{d}{2}(\tau + \log(1 - \tau))}$$

and

$$\mathbb{P}\left(\sum_{i=1}^d Z_i^2 > d(1 + \tau)\right) \leq e^{-\frac{d}{2}(\tau - \log(1 + \tau))}. \quad (15)$$

Lemma 3: Let \mathbf{X} and \mathbf{Y} denote two independent Gaussian vectors of length n with i.i.d. elements. Further, assume that for $i = 1, \dots, n$, $X_i \sim \mathcal{N}(0, 1)$ and $Y_i \sim \mathcal{N}(0, 1)$. Then the distribution of $\mathbf{X}^T \mathbf{Y} = \sum_{i=1}^n X_i Y_i$ is the same as the distribution of $\|\mathbf{X}\|_2 G$, where $G \sim \mathcal{N}(0, 1)$ is independent of $\|\mathbf{X}\|_2$.

B. Calculation of rate-distortion function

In this section we briefly summarize the proof of Example 1 and Example 2.

1) *Proof of Example 1:* For notational simplicity we set $\rho = 1$. Finding a compression algorithm for $\mathcal{B}_2^n(1)$ is equivalent to covering $\mathcal{B}_2^n(1)$ with ℓ_2 -balls of radius D . Consider the following grid points for the interval $[-1, 1]$:

$$\mathcal{G}_1 = \left\{ -\left\lceil \frac{\sqrt{n}}{D} \right\rceil \frac{D}{\sqrt{n}}, \dots, -\frac{D}{\sqrt{n}}, 0, \frac{D}{\sqrt{n}}, \dots, \left\lceil \frac{\sqrt{n}}{D} \right\rceil \frac{D}{\sqrt{n}} \right\}.$$

It is straightforward to show that ℓ_2 -balls of radius D with centers on

$$\mathcal{G}_n = \underbrace{\mathcal{G}_1 \times \mathcal{G}_1 \times \dots \times \mathcal{G}_1}_n.$$

covers the entire space $\mathcal{B}_2^n(1)$. Therefore, our compression scheme maps each vector to its closest codeword, i.e.,

$$\mathcal{D}(\mathcal{E}(\mathbf{x})) = \arg \min_{\mathbf{z} \in \mathcal{G}_n} \|\mathbf{z} - \mathbf{x}\|_2.$$

If the minimizer is not unique, the compression algorithm chooses one of the minimizers at random. The rate such compression algorithm achieves is equal to

$$\begin{aligned} R(D) &= \log \left(2 \left\lceil \frac{\sqrt{n}}{D} \right\rceil + 1 \right)^n \\ &\leq \log \left(2 \frac{\sqrt{n}}{D} + 3 \right)^n \\ &\leq \log \left(5 \frac{\sqrt{n}}{D} \right)^n \\ &= n \log(\sqrt{n}) + n \log \left(\frac{1}{D} \right) + n \log(5). \end{aligned}$$

2) *Proof of Example 2:* Our encoding scheme is inspired by the previous example. The space of all k -sparse signals has $\binom{n}{k}$ hyperplanes. Once we specify the hyperplane \mathcal{H} , $\mathcal{H} \cap \mathcal{B}_2^n(1)$ is an ℓ_2 -ball of radius 1 in k -dimensional subspace. Therefore, according to Example 1 we require

$$\log \left(\frac{\sqrt{k}}{D} \right)^k + ck$$

bits to code it with distortion smaller than D in a specified subspace. Therefore, overall we require $\log \binom{n}{k}$ for coding the subspace, and $\log \left(\frac{2\sqrt{k}}{D} \right)^k$ for specifying the codeword on each hyperplane. This proves

$$R(D) = \log \binom{n}{k} + \log \left(\frac{\sqrt{k}}{D} \right)^k + ck.$$

C. Noiseless measurements

Proof of Theorem 2: Let $\mathbf{x}_o \in \mathcal{Q}$, $\mathbf{y}_o = A\mathbf{x}_o$, $\tilde{\mathbf{x}}_o = \mathcal{D}(\mathcal{E}(\mathbf{x}_o))$, and $\hat{\mathbf{x}}_o = \arg \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{y}_o - A\mathbf{c}\|_2^2$.

Since $\hat{\mathbf{x}}_o$ minimizes the measurement noise $\|\mathbf{y}_o - A\mathbf{c}\|_2^2$ over all $\mathbf{c} \in \mathcal{C}$, we have

$$\begin{aligned} \|\mathbf{y}_o - A\hat{\mathbf{x}}_o\|_2 &= \|A\mathbf{x}_o - A\hat{\mathbf{x}}_o\|_2 \\ &\leq \|A\mathbf{x}_o - A\tilde{\mathbf{x}}_o\|_2 \\ &= \|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2 \|A\mathbf{u}_1\|_2, \end{aligned} \quad (16)$$

where

$$\mathbf{u}_1 \triangleq \frac{A(\mathbf{x}_o - \tilde{\mathbf{x}}_o)}{\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2}.$$

Since the entries of A are i.i.d. Gaussian, \mathbf{u}_1 is a vector of d independent zero-mean Gaussian random variables with variance 1. For $\tau_1 > 0$, define event

$$\mathcal{E}_1 \triangleq \{\|\mathbf{u}_1\|_2^2 < d(1 + \tau_1)\}.$$

By Lemma 2,

$$P(\mathcal{E}_1^c) = P(\|\mathbf{u}_1\|_2^2 \geq d(1 + \tau_1)) \leq e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))}.$$

Since $\tilde{\mathbf{x}}_o$ is the reconstruction of \mathbf{x}_o using the rate-distortion code $(\mathcal{E}, \mathcal{D})$, it follows that $\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2 \leq D$. Therefore, conditioned on \mathcal{E}_1 ,

$$\|\mathbf{y}_o - A\tilde{\mathbf{x}}_o\|_2 \leq D\sqrt{d(1 + \tau_1)}. \quad (17)$$

To find a lower bound on $\|\mathbf{y}_o - A\tilde{\mathbf{x}}_o\|_2$, note that for a fixed $\mathbf{c} \in \mathcal{C}$,

$$\begin{aligned} \|\mathbf{y}_o - A\mathbf{c}\|_2 &= \|A(\mathbf{x}_o - \mathbf{c})\|_2 \\ &= \|\mathbf{x}_o - \mathbf{c}\|_2 \|\mathbf{u}_2\|_2, \end{aligned}$$

where

$$\mathbf{u}_2 \triangleq \frac{A(\mathbf{x}_o - \mathbf{c})}{\|\mathbf{x}_o - \mathbf{c}\|_2}.$$

Similar to \mathbf{u}_1 , \mathbf{u}_2 is a d -dimensional distributed as $\mathcal{N}(\mathbf{0}, I_d)$. Note that \mathbf{u}_2 depends on \mathbf{c} . For $\tau_2 \in (0, 1)$, define event \mathcal{E}_2 as

$$\mathcal{E}_2 \triangleq \{\forall \mathbf{c} \in \mathcal{C} : \|\mathbf{u}_2\|_2^2 \geq d(1 - \tau_2)\}.$$

By Lemma 2 and the union bound, it follows that

$$\begin{aligned} P(\mathcal{E}_2^c) &\leq \sum_{\mathbf{c} \in \mathcal{C}} \{ \|\mathbf{u}_2\|_2^2 \leq d(1 - \tau_2) \} \\ &\leq 2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))}. \end{aligned} \quad (18)$$

Combining the two events, conditioned on that \mathcal{E}_1 and \mathcal{E}_2 both hold,

$$\sqrt{d(1 - \tau_2)}\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \leq D\sqrt{d(1 + \tau_1)}, \quad (19)$$

or equivalently,

$$\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \leq D\sqrt{\frac{1 + \tau_1}{1 - \tau_2}}. \quad (20)$$

Finally, by the union bound,

$$\begin{aligned} P(\mathcal{E}_1 \cap \mathcal{E}_2) &= 1 - P(\mathcal{E}_1^c \cup \mathcal{E}_2^c) \\ &\geq 1 - 2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))} - e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))}. \end{aligned}$$

■

D. Noisy measurements

Proof of Theorem 3: Let $\tilde{\mathbf{x}}_o = \mathcal{D}(\mathcal{E}(\mathbf{x}_o))$. Since by assumption the code operates at distortion D , we have $\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2 \leq D$. On the other hand, since $\hat{\mathbf{x}}_o$ is the solution of (3),

$$\begin{aligned} \|\mathbf{y}_o - A\hat{\mathbf{x}}_o\|_2 &= \|A(\mathbf{x}_o - \hat{\mathbf{x}}_o) + \mathbf{z}\|_2 \\ &\leq \|A(\mathbf{x}_o - \tilde{\mathbf{x}}_o) + \mathbf{z}\|_2. \end{aligned} \quad (21)$$

Expanding both sides of (21) and canceling the common terms, we obtain

$$\begin{aligned} &\|A(\mathbf{x}_o - \hat{\mathbf{x}}_o)\|_2^2 + 2\mathbf{z}^T A(\mathbf{x}_o - \hat{\mathbf{x}}_o) \\ &\leq \|A(\mathbf{x}_o - \tilde{\mathbf{x}}_o)\|_2^2 + 2\mathbf{z}^T A(\mathbf{x}_o - \tilde{\mathbf{x}}_o). \end{aligned} \quad (22)$$

Let

$$\mathbf{u}_1 \triangleq \frac{A(\mathbf{x}_o - \tilde{\mathbf{x}}_o)}{\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2},$$

and

$$\mathbf{u}_2 \triangleq \frac{A(\mathbf{x}_o - \hat{\mathbf{x}}_o)}{\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2}.$$

Using this definition, along with triangle inequality and $a < |a|$, we rewrite (22) as

$$\begin{aligned} &\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2^2 \|\mathbf{u}_2\|_2^2 - 2\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 |\mathbf{z}^T \mathbf{u}_2| \\ &\leq \|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2^2 \|\mathbf{u}_1\|_2^2 + 2\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2 |\mathbf{z}^T \mathbf{u}_1|. \end{aligned} \quad (23)$$

For $\tau_1 > 0$ and $\tau_2 \in (0, 1)$, define events \mathcal{E}_1 and \mathcal{E}_2 as

$$\mathcal{E}_1 \triangleq \left\{ \|A(\mathbf{x}_o - \tilde{\mathbf{x}}_o)\|_2^2 \leq (1 + \tau_1)d\|\mathbf{x}_o - \tilde{\mathbf{x}}_o\|_2^2 \right\}.$$

and

$$\mathcal{E}_2 \triangleq \left\{ \forall \mathbf{c} \in \mathcal{C} : \|A(\mathbf{x}_o - \mathbf{c})\|_2^2 > d(1 - \tau_2)\|\mathbf{x}_o - \mathbf{c}\|_2^2 \right\}.$$

Conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, we can upper bound $\|\mathbf{u}_1\|_2$ by $\sqrt{d(1 + \tau_1)}$, and lower bound $\|\mathbf{u}_2\|_2$ by $\sqrt{d(1 - \tau_2)}$.

In order to bound $|\mathbf{z}^T \mathbf{u}_1|$ and $|\mathbf{z}^T \mathbf{u}_2|$, we employ Lemma 3. Given \mathbf{x}_o , $\tilde{\mathbf{x}}_o$ and $\hat{\mathbf{x}}_o$, both \mathbf{u}_1 and \mathbf{u}_2 are i.i.d. Gaussian vectors with mean zero and variance one, and are both independent of \mathbf{z} . Therefore, by Lemma 3, $\mathbf{z}^T \mathbf{u}_1$ and $\mathbf{z}^T \mathbf{u}_2$ are respectively distributed as $\|\mathbf{z}\|_2 G_1$ and $\|\mathbf{z}\|_2 G_2$, where G_1 and G_2 are zero-mean variance one Gaussian random variables independent of $\|\mathbf{z}\|_2$. For $\gamma_1, \gamma_2 > 1$, define events \mathcal{E}_3 and \mathcal{E}_4 as:

$$\mathcal{E}_3 \triangleq \{|\mathbf{z}^T \mathbf{u}_1| \leq \gamma_1 \sigma \sqrt{d}\},$$

and

$$\mathcal{E}_4 \triangleq \{|\mathbf{z}^T \mathbf{u}_2| \leq \gamma_2 \sigma \sqrt{d}\}.$$

As argued above,

$$\begin{aligned} P(\mathcal{E}_3^c) &= P(\|\mathbf{z}\|_2 |G_1| \geq \gamma_1 \sigma \sqrt{d}) \\ &= P(\|\mathbf{z}\|_2 |G_1| \geq \gamma_1 \sigma \sqrt{d}, \|\mathbf{z}\|_2 \geq \sigma \sqrt{d(1 + \tau_3)}) \\ &\quad + P(\|\mathbf{z}\|_2 |G_1| \geq \gamma_1 \sigma \sqrt{d}, \|\mathbf{z}\|_2 < \sigma \sqrt{d(1 + \tau_3)}) \\ &\leq P(\|\mathbf{z}\|_2 \geq \sigma \sqrt{d(1 + \tau_3)}) + P(|G_1| \geq \gamma_1(1 + \tau_3)^{-0.5}) \\ &\leq e^{-\frac{d}{2}(\tau_3 - \log(1 + \tau_3))} + e^{-\gamma_1^2/2(1 + \tau_3)}. \end{aligned} \quad (24)$$

where $\tau_3 > 0$, and the last line follows from Lemma 2. Adding the union bound to this analysis, we get

$$P(\mathcal{E}_4^c) \leq 2^R \left(e^{-\frac{d}{2}(\tau_4 - \log(1 + \tau_4))} + e^{-\gamma_2^2/2(1 + \tau_4)} \right), \quad (25)$$

where $\tau_4 > 0$.

Conditioned on $\mathcal{E}_1 \cap \dots \cap \mathcal{E}_4$, it follows from (23) that

$$\begin{aligned} &\sqrt{d(1 - \tau_2)}\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2^2 - 2\gamma_2 \sigma \|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \\ &\quad - D^2(1 + \tau_1)\sqrt{d} - 2D\gamma_1 \sigma \leq 0. \end{aligned} \quad (26)$$

Before, analyzing the roots of (26), consider the probability

of $\mathcal{E}_1 \cap \dots \cap \mathcal{E}_4$. By the union bound,

$$P(\mathcal{E}_1 \cap \dots \cap \mathcal{E}_4) \geq 1 - P(\mathcal{E}_1^c) - \dots - P(\mathcal{E}_4^c).$$

To make sure that $P(\mathcal{E}_1 \cap \dots \cap \mathcal{E}_4)$ is close to one for large values of d , it suffices to show that $P(\mathcal{E}_i^c) \rightarrow 0$, for $i = 1, \dots, 4$. By Lemma 2,

$$P(\mathcal{E}_1^c) \leq e^{-\frac{d}{2}(\tau_1 - \log(1 + \tau_1))},$$

and by Lemma 2 and the union bound,

$$P(\mathcal{E}_2^c) \leq 2^R e^{\frac{d}{2}(\tau_2 + \log(1 - \tau_2))}.$$

Upper bounds on $P(\mathcal{E}_3^c)$ and $P(\mathcal{E}_4^c)$ are given in (24) and (25), respectively. Choose $\tau_1 = \tau_3 = 1$, $\tau_2 = 1 - D$, $\tau_4 = -\log(eD)$,

$$\gamma_1 = \sqrt{\frac{4R}{\log \frac{1}{eD}}},$$

and

$$\gamma_2 = \sqrt{4R \log \frac{1}{eD}}$$

For $\tau_1 = 1$,

$$P(\mathcal{E}_1^c) \leq e^{-\frac{0.6\eta R}{\log 1/(eD)}},$$

Similarly, for $\tau_3 = 1$, from (24),

$$P(\mathcal{E}_3^c) \leq e^{-\frac{0.6\eta R}{\log 1/(eD)}} + e^{-\frac{R}{\log 1/(eD)}}. \quad (27)$$

Following an analysis similar to one detailed in (5) yields

$$\begin{aligned} R \log 2 + \frac{d}{2}(\tau_2 + \log(1 - \tau_2)) \\ \leq (1 - 2\eta)R, \end{aligned} \quad (28)$$

and therefore,

$$P(\mathcal{E}_2^c) \leq e^{-(2\eta - 1)R}.$$

The remaining probability that we need to bound is $P(\mathcal{E}_4^c)$. For $\tau_4 = -\log(eD)$,

$$\begin{aligned} R \log 2 - \frac{d}{2}(\tau_4 - \log(1 + \tau_4)) \\ \leq R \log 2 - \frac{2\eta R}{\log \frac{1}{eD}}(\log \frac{1}{eD} - \log(1 + \log \frac{1}{eD})) \\ \leq (\log 2 - 2\eta)R + \frac{2\eta R}{\log \frac{1}{eD}} \log(1 + \log \frac{1}{eD}) \\ \leq (\log 2 - 0.8\eta)R, \end{aligned} \quad (29)$$

where the last line holds because by assumption, $D < (5e)^{-1}$, or $(De)^{-1} > 5$, and for $t > 5$, $\log(1 + \log(t))/\log t < 0.6$. Also, since $\log t/(1 + \log t) > 0.5$, for $t > 5$,

$$\begin{aligned} R \log 2 - \frac{\gamma_2^2}{2(1 + \tau_4)} &= R \log 2 - \frac{2R \log \frac{1}{eD}}{(1 + \log \frac{1}{eD})} \\ &< R(\log 2 - 1) < -0.3R, \end{aligned} \quad (30)$$

Therefore,

$$P(\mathcal{E}_4^c) \leq e^{-(0.8\eta - \log 2)R} + e^{-0.3R}. \quad (31)$$

Going back to the quadratic equation (26) and inserting the

parameters values, we derive

$$\begin{aligned} D \sqrt{\frac{4\eta R}{\log \frac{1}{eD}}} \|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2^2 - 2\sigma \sqrt{4R \log \frac{1}{eD}} \|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2 \\ - 2D^2 \sqrt{\frac{4\eta R}{\log \frac{1}{eD}}} - 2D\sigma \sqrt{\frac{4R}{\log \frac{1}{eD}}} \leq 0. \end{aligned} \quad (32)$$

The quadratic equation $ax^2 - 2bx - c = 0$, with $a, b, c > 0$, has one positive and one negative root. Hence, $\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2$ is smaller than the positive root of (32). Therefore, conditioned on $\mathcal{E}_1 \cap \dots \cap \mathcal{E}_4$, from (32), $\|\mathbf{x}_o - \hat{\mathbf{x}}_o\|_2$ is upper bounded by

$$\frac{\sigma}{D\sqrt{\eta}} \log \frac{1}{eD} + \sqrt{2D + \frac{2\sigma}{\sqrt{\eta}} + \frac{\sigma^2}{\eta D^2} \log^2(\frac{1}{eD})}. \quad (33)$$

■

E. Proof of Theorem 4

Lemma 4: If $f \in L_2^r([0, 1])$, then the distribution of $\int_0^1 f(t) dW$ is $N(0, \|f\|_2^2)$.

Proof: For notational simplicity we consider a continuous function $f > 0$. Extension to piecewise continuous functions with finite number of discontinuities is straightforward. Note that since the functions are defined on $[0, 1]$, according to Heine-Cantor theorem, they are also uniformly continuous. For partition \mathcal{P} , $0 = t_0 < t_1 < \dots < t_{n-1} < t_n = 1$, define the partition length as

$$\delta = \sup_i |t_i - t_{i-1}|.$$

Consider a sequence of partitions $\{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_n, \dots\}$ such that

$$\delta_{\mathcal{P}_n} \leq \frac{1}{n}.$$

Given the division points $0 = t_{n,0} < t_{n,1} < t_{n,2} < \dots < t_{n,q_n} = 1$, define the following piecewise constant function:

$$f_n(t) = \sum_{i=1}^{q_n} \left(\min_{t \in [t_{i-1}, t_i]} f(t) \right) 1_{\{t_{i-1} \leq t \leq t_i\}}.$$

It is clear from the uniform continuity that for every $\epsilon > 0$ there exists n_ϵ such that for every $n > n_\epsilon$

$$|f(t) - f_n(t)| < \epsilon \quad \text{for } t \in [0, 1].$$

Therefore, for $n > n_\epsilon$

$$\int_0^1 (f(t) - f_n(t))^2 dt \leq \epsilon.$$

In other words,

$$\lim_{n \rightarrow \infty} \int_0^1 (f(t) - f_n(t))^2 dt = 0.$$

Now that we have constructed a sequence of simple functions, we have

$$\int f dW = \lim_{n \rightarrow \infty} \int f_n dW.$$

It is straightforward to show that

$$\int f_n dW = \sum_{i=1}^{q_n} \left(\min_{t \in [t_{i-1}, t_i]} f(t) \right) (W(t_i) - W(t_{i-1})).$$

Which has a Gaussian distribution with mean zero and variance $\sum_{i=1}^{q_n} (\min_{t \in [t_{i-1}, t_i]} f(t))^2 (t_i - t_{i-1})$. As $n \rightarrow \infty$ this sequence of Gaussian random variables converge to $N(0, \|f\|^2)$. On the other hand according to Ito's integral definition they converge in mean square to $\int f dW$. Therefore, $\int f dW$ is distributed as $N(0, \|f\|_2^2)$. ■

Lemma 5: Let $f \in L_2^r([0, 1])$ and consider

$$\mathcal{A}(f) = \left[\int_t f dW_1(t), \dots, \int_t f dW_d(t) \right],$$

where W_1, W_2, \dots, W_d are independent Brownian motions. Then $\frac{\|\mathcal{A}(f)\|_2^2}{\|f\|_2^2}$ is a χ^2 random variable with d degrees of freedom.

Proof: According to Lemma 4 and independence of W_i s, the elements of $\mathcal{A}(f)$ are iid $N(0, \|f\|_2^2)$. Therefore $\mathcal{A}(f)/\|f\|_2$ is iid $N(0, 1)$ vector, that proves

$$\frac{\|\mathcal{A}(f)\|_2^2}{\|f\|_2^2} \sim \chi^2(d).$$

Proof: Given Lemmas 5 and 4 the proof is essentially the same as the proof of Theorem 2. Therefore, we briefly mention the main steps. Let $\tilde{f}_o = \mathcal{D}(\mathcal{E}(f))$ and \hat{f} denote the reconstruction of CSP. Since \hat{f} is the solution of CSP, we have

$$\|\mathcal{A}(\hat{f}) - \mathcal{A}(f_o)\|_2^2 \leq \|\mathcal{A}(\tilde{f}) - \mathcal{A}(f_o)\|_2^2$$

According to Lemma 5 both $\|\mathcal{A}(\tilde{f}) - \mathcal{A}(f_o)\|_2^2 / \|\tilde{f} - f_o\|_2^2$ and $\|\mathcal{A}(\hat{f}) - \mathcal{A}(f_o)\|_2^2 / \|\hat{f} - f_o\|_2^2$ are χ^2 random variables with d degrees of freedom. Therefore it is straightforward to confirm that

$$P(\|\mathcal{A}(\tilde{f}) - \mathcal{A}(f_o)\|_2 \geq D\sqrt{d(1+\tau)}) \leq e^{-\frac{d}{2}(\tau - \log(1+\tau))},$$

and also for every $\hat{f} \in \mathcal{C}_R$

$$\|\mathcal{A}(\hat{f}) - \mathcal{A}(f_o)\|_2 \geq \|f_o - \hat{f}_o\|_2 \sqrt{d(1-\tau)}.$$

with probability $1 - 2R_e d/2(\tau + \log(1-\tau))$. Combining these results completes the proof. ■

F. Proof of Theorem 6

Let $\mathbf{x}_o \in \mathcal{Q}$, $\mathbf{y}_o = \mathbf{A}\mathbf{x}_o$, $\tilde{\mathbf{x}}_o = \mathcal{D}(\mathcal{E}(\mathbf{x}_o))$, and $\hat{\mathbf{x}}_o = \arg \min_{\mathbf{c} \in \mathcal{C}} \|\mathbf{y}_o - \mathbf{A}\mathbf{c}\|_2^2$. As before, we have $\|\mathbf{y}_o - \mathbf{A}\hat{\mathbf{x}}_o\|_2 \leq \|\mathbf{y}_o - \mathbf{A}\tilde{\mathbf{x}}_o\|_2$. Hence,

$$\|\mathbf{A}\tilde{\mathbf{x}}_o - \mathbf{A}\hat{\mathbf{x}}_o\|_2 - \|\mathbf{A}\tilde{\mathbf{x}}_o - \mathbf{A}\mathbf{x}_o\|_2 \leq \|\mathbf{A}\mathbf{x}_o - \mathbf{A}\tilde{\mathbf{x}}_o\|_2.$$

Rearranging the terms proves that

$$\|\mathbf{A}\tilde{\mathbf{x}}_o - \mathbf{A}\hat{\mathbf{x}}_o\|_2 \leq 2\|\mathbf{A}\mathbf{x}_o - \mathbf{A}\tilde{\mathbf{x}}_o\|_2 \leq 2\sigma_{\max}(A)D, \quad (34)$$

where $\sigma_{\max}(A)$ is the maximum singular value of A . Define

$$\mathcal{T} = \{\mathbf{x}_1 - \mathbf{x}_2 \mid \mathbf{x}_1 \in \mathcal{C}_R, \mathbf{x}_2 \in \mathcal{C}_R\}.$$

Define the event $\mathcal{E}_1^{(n)}$ as

$$\mathcal{E}_1 \triangleq \{\nexists \mathbf{h} \in \mathcal{T}; \|\mathbf{A}(\mathbf{h})\|_2 < \tau\sqrt{d}\|\mathbf{h}\|_2\}, \quad (35)$$

and, for $t > 0$, the event $\mathcal{E}_2^{(n)}$ as

$$\mathcal{E}_2 \triangleq \left\{ \sigma_{\max}(A) - \sqrt{d} - \sqrt{n} < t\sqrt{d} \right\}. \quad (36)$$

Using the union bound and Lemmas 5, 2 it is straightforward to confirm that

$$P(\mathcal{E}_1^c) \leq 2^{2R_e} e^{\frac{d}{2}(\tau + \log(1-\tau))} \quad (37)$$

Finally, using the results on the concentration of Lipschitz functions of a Gaussian random vector [54], we obtain

$$\begin{aligned} P(\mathcal{E}_2^c) &= P\left(\sigma_{\max}(A) - \sqrt{d} - \sqrt{n} > t\sqrt{d}\right) \\ &\leq e^{-dt^2/2}. \end{aligned} \quad (38)$$

Combining (37) and (38) with (34) finishes the proof.

VII. CONCLUSION

In this paper, we studied the problem of employing a family of compression algorithms for compressed sensing, i.e., recovering structured signals from their undersampled set of random linear measurements. Addressing this problem enables CS schemes to exploit complicated structures integrated in compression algorithms. We proposed compressible signal pursuit (CSP) algorithm that outputs the codeword that best matches the measurements. We proved that employing a family of compression algorithms whose rate-distortion function satisfies $\limsup_{D \rightarrow 0} R(D)/\log(1/D) \leq \alpha$, with α smaller than the ambient dimension, with high probability, CSP recovers signals from 4α measurements. CSP is also applicable to infinite-dimensional signal classes. The CSP algorithm is still computationally demanding and requires approximation or simplification for practical applications. This important direction is left for future research.

REFERENCES

- [1] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, Apr. 2006.
- [2] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, Feb. 2006.
- [3] S. Bakin. Adaptive regression and model selection in data mining problems. *Ph.D. Thesis, Australian National University*, 1999.
- [4] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Processing*, 58(6):3042–3054, Jun. 2010.
- [5] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67, 2006.
- [6] S. Ji, D. Dunson, and L. Carin. Multi-task compressive sensing. *IEEE Trans. Signal Processing*, 57(1):92–106, 2009.
- [7] A. Maleki, L. Anitori, Z. Yang, and R. G. Baraniuk. Asymptotic analysis of complex lasso via complex approximate message passing (CAMP). *arXiv:1108.0477v1*, 2011.
- [8] M. Stojnic. Block-length dependent thresholds in block-sparse compressed sensing. *Preprint arXiv:0907.3679*, 2009.
- [9] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Processing*, 57(8):3075–3085, 2009.
- [10] M. Stojnic. ℓ_2/ℓ_1 -optimization in block-sparse compressed sensing and its strong thresholds. *IEEE J. Select. Top. Signal Processing*, 4(2):350–357, 2010.

- [11] L. Meier, S. Van De Geer, and P. Bühlmann. The group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, 70(Part 1):53–71, 2008.
- [12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Preprint*, 2010.
- [13] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, Apr. 2010.
- [14] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, Apr. 2010.
- [15] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Processing*, 50(6):1417–1428, Jun. 2002.
- [16] S. Som and P. Schniter. Compressive imaging using approximate message passing and a markov-tree prior. *IEEE Trans. Signal Processing*, 60(7):3439–3448, 2012.
- [17] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *arXiv preprint arXiv:0912.3599*, 2009.
- [18] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optimization*, 21(2):572–596, 2011.
- [19] M. F. Duarte, W. U. Bajwa, and R. Calderbank. The performance of group Lasso for linear regression of grouped variables. Technical report, Technical Report TR-2010-10, Duke University, Dept. Computer Science, Durham, NC, 2011.
- [20] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2002.
- [21] A. Maleki, M. Shahram, and G. Carlsson. A near optimal coder for image geometry with adaptive partitioning. In *IEEE Int. Conf. Image Processing*, pages 1061–1064, 2008.
- [22] A.N. Kolmogorov and V.M. Tikhomirov. ε -entropy and ε -capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.
- [23] S. Sarvotham, D. Baron, and R.G. Baraniuk. Measurements vs. bits: Compressed sensing meets information theory. In *Proc. Allerton Conf. Communication, Control, and Computing*, 2006.
- [24] V.K. Goyal, A.K. Fletcher, and S. Rangan. Compressive sampling and lossy compression. *Signal Processing Magazine*, 25(2):48–56, 2008.
- [25] P. Boufounos and R. Baraniuk. Quantization of sparse representations. Technical report, Rice University Technical Report, 2007.
- [26] E. Candès and J. Romberg. Encoding the ℓ_p ball from limited measurements. In *Proc. Data Compression Conference (DCC)*, pages 33–42. IEEE, 2006.
- [27] W. Dai, H.V. Pham, and O. Milenkovic. Distortion-rate functions for quantized compressive sensing. In *Proc. Information Theory Workshop (ITW)*, pages 171–175, 2009.
- [28] J.Z. Sun and V.K. Goyal. Optimal quantization of random measurements in compressed sensing. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 6–10. IEEE, 2009.
- [29] C. Deng, W. Lin, B. Lee, and C.T. Lau. Robust image compression based on compressive sensing. In *Proc. Int. Conference Multimedia and Expo (ICME)*, pages 462–467, 2010.
- [30] A. Schulz, L. Velho, and E.A.B. Da Silva. On the empirical rate-distortion performance of compressive sensing. In *Proc. Int. Conference Image Processing (ICIP)*, pages 3049–3052, 2009.
- [31] P.T. Boufounos. Universal rate-efficient scalar quantization. *IEEE Trans. Info. Theory*, 58(3):1861–1872, 2012.
- [32] K. Falconer. *Fractal geometry: mathematical foundations and applications*. Wiley, 2003.
- [33] J. Shihao, X. Ya, and L. Carin. Bayesian compressive sensing. *IEEE Trans. Signal Processing*, 56(6):2346–2356, Jun. 2008.
- [34] P. Schniter. Turbo reconstruction of structured sparse signals. In *Proc. IEEE Conf. Inform. Science and Systems (CISS)*, Mar. 2010.
- [35] D. Baron, D. Guo, and S. Shamai. A single-letter characterization of optimal noisy compressed sensing. Sep. 2009.
- [36] S. Rangan, A. K. Fletcher, and V. K. Goyal. Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing. *arXiv preprint arXiv:0906.3234v3*, 2009.
- [37] D. L. Donoho, A. Maleki, and A. Montanari. Noise sensitivity phase transition. *IEEE Trans. Inform. Theory*, 57(10), Oct. 2011.
- [38] D. L. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, 2009.
- [39] D. L. Donoho and J. Tanner. Precise undersampling theorems. *Proc. IEEE*, 98(6):913–924, Jun. 2010.
- [40] Y. Wu and S. Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inform. Theory*, 56(8):3721–3748, Aug. 2010.
- [41] A. Maleki and D. L. Donoho. Optimally tuned iterative thresholding algorithm for compressed sensing. *IEEE J. Select. Top. Signal Processing*, Apr. 2010.
- [42] Y. Wu and S. Verdú. Optimal phase transitions in compressed sensing. *IEEE Trans. Info. Theory*, 58(10):6241–6263, oct. 2012.
- [43] D.L. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. In *Proc. Int. Symposium Info. Theory (ISIT)*, pages 1231–1235, 2012.
- [44] T. Kawabata and A. Dembo. The rate-distortion dimension of sets and measures. *IEEE Trans. Info. Theory*, 40(5):1564–1572, 1994.
- [45] S. Jalali, A. Maleki, and R. Baraniuk. Minimum complexity pursuit for universal compressed sensing. *arXiv preprint arXiv:1208.5814*, 2012.
- [46] D. L. Donoho. The Kolmogorov sampler, Jan 2002.
- [47] D. L. Donoho, H. Kakavand, and J. Mammen. The simplest solution to an underdetermined system of linear equations. In *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, pages 1924–1928, Jul. 2006.
- [48] D. Baron and M. F. Duarte. Signal recovery in compressed sensing via universal priors. *Arxiv preprint arXiv:1204.2611*, 2012.
- [49] D. Baron and M. F. Duarte. Universal MAP estimation in compressed sensing. Sep. 2011.
- [50] S. Jalali and A. Maleki. Minimum complexity pursuit. In *Proc. Allerton Conf. Communication, Control, and Computing*, pages 1764–1770, 2011.
- [51] S. Jalali, A. Maleki, and R. Baraniuk. Minimum complexity pursuit: Stability analysis. In *Proc. Int. Symposium Info. Theory (ISIT)*, pages 1857–1861, 2012.
- [52] S. Jalali and T. Weissman. Rate-distortion via markov chain monte carlo. In *Proc. IEEE Int. Symp. Info. Theory (ISIT)*, pages 852–856, 2008.
- [53] S. Jalali, A. Montanari, and T. Weissman. Lossy compression of discrete sources via the viterbi algorithm. *IEEE Trans Info. Theory*, 58(4):2475–2489, 2012.
- [54] E. Candès, J. Romberg, and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, Dec. 2005.